Complex Intell. Syst. (2017) 3:133–151
DOI 10.1007/s40747-017-0040-1

ORIGINAL ARTICLE

Big data for secure healthcare system: a conceptual design

Bikash Kanti Sarkar[1]

Received: 7 October 2016 / Accepted: 8 March 2017 / Published online: 21 March 2017

Abstract The concept of big data is now treated from different points of view covering its implications in many fields remarkably including healthcare. To achieve the wealth of health information, integrating, sharing and availing data are the essential tasks that ultimately demand the concept of distributed system. However, *privacy* and *security* of data are the matter of concern, as data need to be accessed from various locations in the distributed system. The present study first provides a broad overview on big data and the effectiveness of healthcare big data for non-expert readers. Then, this article builds a distributed framework of organized healthcare model for the purpose of protecting patient data.

Keywords Big data · Healthcare · Framework · Privacy · Security

Introduction

The concept of '*big data*' is not new; however, the way it is defined is constantly changing. In practice, a data set is considered '*big*' if it ranges from a few terabytes (1 TB = 2^{40} bytes) to many petabytes (1PB = 2^{50} bytes) but the term *big data* technically implies that the generation rate is unprecedented. The statistical estimation roughly reports that 90% of the current data is created in the last couple of years (http://www-01.ibm.com/software/data/bigdata/). According to Bakshi and Kapil [1], the size of digital data in 2011 is roughly 1.8 Zettabytes (1.8×10^{21} bytes) and they assert that the supporting network infrastructure has to manage 50 times more information by year 2020. Obviously, it makes real *concern* over storing data as well as processing them. But huge amounts of data with variation assist to design effective predictive model. In fact, decisions that were previously based on guesswork can now be made based on the available data itself. So, we were unlucky earlier in this sense.

Undoubtedly, the phrase 'big data' has now become very popular to describe precisely the exponential growth and availability of data, in both structured and unstructured way. In particular, the term itself was first introduced by Roger Magoulas from O'Reilly media in 2005 to define great amount of data which cannot be managed and processed by traditional data management techniques (due to the complexity and size of the data). It is interesting to note here that a study [2] claims that the term was found in 1970s but it has been first comprised in the publication of 2008. Anyway, the present form of big data defines data by its size, comprising large, complex, and independent collection of data sets. Also, people generally agree that big data should have four standard characteristics(called 4V's namely volume, variety, velocity, and veracity) as suggested by IBM. Each of these four V's is explained in Sect. 2.

At the present date, big data and its analytics are being effectively used in many fields, e.g., *Information Technology* improvises the scope of improvement in security troubleshooting, *customer care service* enhances the customer satisfaction based on the identification of customer patterns, *online transaction* assists to detect *fraud*, *risk management* in business and commerce forecasts a bigger picture in risk factors, *astronomy* helps to know more about universe, healthcare system provides us quality services, and so on.

In healthcare system, the information stored in health database has enhanced over the past ten years, leading it to be considered big data. According to Raghupathi, this industry

✉ Bikash Kanti Sarkar
bk_sarkarbit@hotmail.com

[1] Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, India

has historically generated huge amounts of data driven by record keeping and patient care [3]. This massive quantity of data hold the promise of supporting a wide range of medical and healthcare functions, including clinical decision support, sensor-based health condition and food safety monitoring, disease surveillance, and population health management, etc. [4–6]. *For instance*, diagnosing cancer requires petabytes of data from various sources to identify the state of the disease and the survival potential of the patient. Further, the use of information technology on healthcare big data today is reducing the *cost of healthcare* while improving its quality by emphasizing more preventive and personalized care and basing on continuous monitoring [7]. In this context, James et al. give an estimation of savings $300 Billion every year in the US alone [8].

However, to satisfy the above-mentioned health services, health information needs to be accessible and available to everyone involved in the healthcare system. In this respect, a study [9] suggests that a high-level integration of data, interoperability, and its sharing are essential among different healthcare practitioners and institutions to deliver secured high-quality healthcare to the patients they serve. In 2011, Kuo reported that cloud computing (a form of distributed computing) is a developing phenomenon in the field, Information Communication Technology (ICT) field that has gained increasing attention from healthcare organisations to overcome some of the e-health barriers [10]. He claimed that ICT can offer economic savings by decreasing the initial and operational costs of e-health to a great extent. In 2014, Sultan asserted that cloud technology in context of healthcare means that fewer technicians will be required by the healthcare organisations [11]. Recently, Alharbi et al. [12] identify the factors to influence the adoption of Cloud Computing in Saudi healthcare organisations. Further, Peddi et al. propose an intelligent cloud-based data processing model for mobile e-health multimedia applications [13]. The model mainly focuses on the intelligent central cloud broker for single, mixed, and multiple food object images, proposing dynamic cloud allocation mechanism.

Thus, with the help of big data phenomenon, we are gaining many things like extracting useful patterns, detecting frauds, managing risk factors, reducing healthcare cost and many more, but we are simultaneously facing many challenges such as collection of data, storage of data, data curation, data analysis, data security, etc. Alternately, these are also the research opportunities to us. Again, the review on healthcare system makes clear that the researchers are recently paying attention to healthcare cloud. Fortunately, research on various security issues surrounding healthcare information systems has been heated over the past few years. In particular, ISO/TS 18308 standard gives the definitions of security and privacy issue for electronic health records [14]**.** However, we do not find sufficient studies that focus primarily on designing secure cloud-based healthcare system. Of course, the present article includes a study [15] that has contributed a conceptual framework in this respect. But the framework lacks many things such as broad overview, security, implementation details, etc. in context of cloud-based e-healthcare system.

Contributions of the study

In this article, a broad overview on big data from different aspects is first presented. The study then focuses on healthcare big data and introduces a conceptual distributed healthcare framework extending the idea suggested by Ahmed [15]. The presented model provides high-level implementation details and emphasizes to preserve higher degree of privacy and security of patient *sensitive* data. It is important to note here that healthcare data consist of lots of patient sensitive data, so privacy and protection of individual's sensitive data are essential; otherwise, personal data can be misused without permission.

Organization of the paper

The paper is organized as follows: Sect. 1 introduces big data and its importance. The impact of healthcare data and its analytics are highlighted in this section through several literature reviews. Section 2, describes about big data, data characteristics and some of its major aspects. This also gives a review on the state-of-the-art distributed processing framework for managing big data. Section 3 primarily discusses healthcare information, its opportunities, issues, and challenges. Section 4 sates the needs of simple e-health system and the distributed e-health system are discussed in this section too. Section 5 briefly discusses the conceptual e-health framework introduced by Ahmed E. Youssef, whereas the proposed cloud-based healthcare architecture comprising a new security model is detailed in Sect. 6. Immediate next section presents an analysis of the proposed system, including its managerial implications. Concluding remarks and future scopes on the study are summarized in Sect. 8.

Big data: an overview

Before embarking on the discussion of healthcare big data, it is necessary to provide a clear picture about big data, covering its characteristics and some of its distinctive points such as sources, comparison with traditional data, opportunities, challenges, etc.

Characteristics of big data

Recall that 'big data' is characterized by V's. In the present section, each of 4 V's as suggested by IBM in 2013 is illustrated below.

Volume (size) This characteristic represents the quantity of data (usually measured in Terabytes to Zettabytes) gathered by organizations from several locations.
Variety (different data formats) It says about data types that big data can comprise. In fact, it is composed of text, image, video, audio or other forms of data. Hence, the term '*variety*' suggests that the data may be *structured* (e.g., Relational data), *unstructured* (e.g., Word, PDF, Text, Media logs, etc.) or *semi-structured* (e.g., XML data, csv: comma separated value). In other words, it refers to heterogeneous data. One may note that structured data are tagged, and can easily be stored as well as analyzed, but unstructured data are scattered and difficult to analyze. On the other hand, semi-structured data do not conform to fixed fields but contains tags to separate data elements [16].
Velocity (speed of data generation) This tells how fast data grow. According to [17], data evolve very rapidly and the generated unprecedented quantity of data needs to be stored, transmitted, and processed quickly, since many activities are very important in real-life and they need immediate responses [17], e.g., detecting infectious as early as possible.
Veracity The term relates to *quality*, *relevance*, *predictive value*, and *meaning* of data. Precisely, this feature ensures the degree of *trusts* to the leader of an organization to make decision. So, establishing the right correlation among these qualities in big data is very important for the business future. From veracity point of view, Leventhal states that big data creates big value [18]. Priyanka and Nagarathna discussed about 'the data being stored and their meaningful mining' [19]. They also commented that paying attention on veracity in *data analysis* is rather a challenging task than managing other characteristics like *volume* and *velocity*.

Conceptually, the first three Vs relate, namely data collection, storage, and transmission of data, and these associate with *data engineering*, whereas the last V deals combiningly with analytics, statistical methods, knowledge extraction, and decision-making and all these come under *data science*. At this point, it is safe to say that some other communities or organizations have expressed their opinions on big data as follows:

- In 2010, Apache Hadoop defined big data as 'the data sets that could not be captured, managed and processed by general computers within an acceptable scope' [20].
- In 2011, an IDC report revealed big data as 'big data technologies describe a new generation of technologies and architectures, designed to extract economical *value* from large*volume* of a wide *variety* of data by enabling the *high-velocity* capture, discovery,and/or analysis' [21]. According to this definition, features of big data may be ultimately summarized as 5Vs, where the first 4 Vs are identical as defined by IBM and a new feature named as *value* (refers to *social value*) is assumed to add to those Vs.
- A report delivered to the U.S. Congress in August 2013 defines big data as—'large volumes of high-velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information' [22].

Thus, the definition of big data changes in accordance with the development of processing power, memory size, data transfer rate on disks and networks, data organization and representation, and analytical models. As an evidence, big data presently expresses itself as 5Vs (instead of 4 V's), including the important property named as *Value* (i.e., long-range values of attributes), whereas it was 3 Vs earlier to 4 Vs. So, it is very difficult to provide a universally accepted definition on big data. Mayer-Schönberger and Cukier [23] argued that there is no rigorous definition of 'big data'. They have noted that, according to http://datascience.berkeley.edu/what-is-big-data/ (accessed on September 14, 2014), there exist at least 43 different definitions for big data.

Some distinctive points of big data

In this section, some distinctive points about big data are briefly stated. For more, one may refer the review articles [24,25]. Let us start with discussing *source*s of big data.

Sources of big data

One may have a *query* about the sources of big data. The very simple and practical answer is, big data surrounds us but we are unable to realize it due to lack of our experience. To get better understanding of the sources, we may take few examples such as sensors, CC TV Camera, Social Network, Online Shopping, Airlines, Weather Forecasting, Banking, Education, etc. After all, healthcare is a prime example of the sources of big data and it is spread among multiple healthcare systems, health insurers, researchers, government entities and so forth.

Big data lifecycle

By definition, the big data lifecycle (BDLC) involves multiple distinct phases as explained below.

Data collection (acquisition) Big data does not arise out of a vacuum. It is recorded from several data-generating sources.

This phase concerns with collection of data from various data sources and stores them in system like Hadoop Distributed File System (HDFS).

Data cleaning Truly, the information collected from various sources may not necessarily be in a format ready for analysis. *For example*, electronic health records captured from hospitals comprise transcribed dictations from several physicians, image data such as X-rays, and structured data from sensors and measurements (possibly with some associated uncertainty). Obviously, the collected information may not be *clear* and *complete*, i.e., they may be *junky* and consist of *missing values*. Besides, there may exist *inconsistency* in data and much of the data may be of no interest. So, it is necessary to filter out and compress them by orders of magnitude. However, we should not discard useful information while filtering data. In fact, deciding the information to be removed or corrected is carried out at this stage, and it is a continuing technical challenge.

Data aggregation and representation Given the heterogeneity of the flood of data, it is not enough to record the data and simply throw it into a repository. In particular, this phase first aggregates the data of different formats and finally represents them into a *common* format. However, any data irrespective of whether structured, semi-structured, and unstructured should be first meaningfully analyzed to accept or reject. Ideally, if the data are purposeful, then we accept it and convert it into the desired format. *For instance*, medical data consist of mostly unstructured data such as hand-written physician notes but these must be processed to check their relevance for acceptance before representation, i.e., not all data may be relevant to store and convert. Especially, the data which are useful, only these are accepted.

Data modelling and analysis This phase of BDLC is also called as big data analytics (BDA). It is one of the fastest evolving fields due to convergence of IoT, the cloud and smart assets [26]. The process, BDA, primarily examines large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations, and other business benefits. Also, BDA is essentially needed by the cloud users because it can utilize huge amounts of data to make faster and better decisions to respond the changes and uncertainties. Certainly, BDA is not just a technology; it is an integral toolset of strategy, marketing, human resources, and research. In particular, it is different in many respects from traditional modelling employed over small samples. As per Davis [27], BDA contains a set of well-established and widely used analytical methodologies and tools such as correlations, cluster analysis, filtering, decision trees, Bayesian analysis, neural network analysis, regression analysis, and textural analysis.

Table 1 Comparison between traditional data and big data

Sl. no.	Traditional data	Big data
(i)	Traditional data are usually measured in GB (gigabyte)	It requires TB or PB for big data measurement
(ii)	The growth of traditional data is measured on hour or day basis	No such period is fixed for big data
(iii)	The format of traditional data is assumed to be structured	Big data may be structured or unstructured or semi-structured
(iv)	Integration of data in context of traditional data is simple	It is rather difficult and time consuming for big data
(v)	In general, RDBMS is used for managing traditional data	Architectures like Hadoop based File System with MapReduce, NoSQL (not only SQL), High Performance Computing System (HPCS) are used for storing and analysing big large data sets reliably
(vi)	Access to traditional data is interactive	Batch or near to real-time system is necessary for managing big data

Concisely, this stage deals with the methods for *querying* and *mining* big data in order to design predictive model for unseen data, and the designed model finally performs analysis on the classified data. *For example*, Government may require a list of *malnourished children* in a location. In this respect, it is first necessary to gather family-wise details of the location. Then, we must identify the children whose family are below poverty line. These data are now processed to generate the health report of children.

Data delivery It involves generation of report based on modelling of data. *For instance*, a report comprising malnourished children at a particular location can be made to take appropriate precaution. So, it helps the government to take necessary measures to avoid any further complications.

Comparison between big data and traditional data

Undoubtedly, there are significant differences between traditional and big data. Table 1 offers a brief comparison between them.

Challenges of big data

Certainly, the emergence of big data has supplied us unprecedented large-scale samples when dealing with computational

problem. Today, this effect causes many challenges in harnessing the potential of big data. Some major challenges are stated below.

Data capture and storage Data sets grow exponentially day-by-day because they are continuously gathering ubiquitous information-sensing mobile devices, aerial sensory technologies, remote sensing, software logs, cameras, microphones, radio-frequency identification readers, wireless sensor networks, and so on. This makes a real challenge in front of us.

Big data analytics (BDA) Recall that BDA is an important part of BDLC. In the recent era, determining the best strategy for data analytics is an important task to an organization. Technically, it is well accepted that 'the larger the data set to be processed, the longer it will take to analyze'. Therefore, it is essential to design a system that effectively operates on voluminous data and shows always better performance. In addition, deciding the frequency and the interval of analysis together is also a crucial job.

Veracity or trustworthiness of data It poses a major challenge with regard to volume, variety, and velocity.

Privacy and security It may be reminded that big data consists of large amount of complex data. Sorting such complex data on the basis of privacy levels and applying security over those are very difficult tasks for any organization. Further, several companies are now-a-days carrying out business across the countries and the continents. Certainly, the differences in privacy laws in such cases need to be taken into consideration while starting the big data initiative.

Need of IT specialists According to the study of James et al. [8] on big data, there is a need of 190,000 or more workers with analytical expertise and 1.5 million more data-literate managers only in the United States. The statistics may considerably assist the organizations to plan either to hire experts or to train existing employees in the new fields. Importantly, it seems to be a great challenge to the organization.

In addition, resolving *heterogeneity, incompleteness, scalability of big data* are some other challenging tasks.

State-of-the-art big data analytical methods

In the present section, a distributed file system, named Hadoop, is first discussed; then it gives an insight into MapReduce (a programming tool) for Hadoop in order to process data.

Hadoop: *a distributed file architecture*
To deal with the challenges of storing and accessing big data, one distributed cluster *platform* is necessary. Such a system must provide large storage space (petabyte) and location transparent access to data files to the servers on the cluster. Hadoop Distributed File System (HDFS) [28] is an example of cluster file system which is designed for reliably storing large amount of various structure or no structure data across machines in a large scale cluster. Interestingly, HDFS was originally derived from Google Files System (GFS) paper [29]. It has ability to deliver an open source cluster file system similar to GFS.

The Hadoop cluster system contains a number of nodes in which one node named NameNode (also called as master node in GFS) is a dedicated node, and the rest are DataNodes (also called as slave nodes). The DataNodes are fully connected and they communicate with each other using TCP-based protocols. In NameNode, the metadata are stored, whereas application data are stored on DataNodes. In particular, the NameNode provides all the necessary services to the DataNodes, whereas the DataNodes mainly do the computing task in parallel. Hence, the Hadoop architecture follows *master-slave* architecture.

In HDFS environment, a data file is split into one or more blocks, and the blocks are then replicated across several DataNodes. When an HDFS client needs to access a file, it first contacts with NameNode to get the locations of data blocks comprising the file and then reads these data blocks from the closest DataNode(s). Hence, Hadoop has the potential to process extremely large volumes of data mainly by allocating partitioned data sets to numerous nodes, each of which has capability to solve different part of the large problem in parallel. Another highlighting feature of Hadoop system is that it has high fault-tolerance capability because if a slave node fails, then the master node will detect it and reassign the work to other slave nodes. However, a highly fault-tolerant distributed file system that is responsible for storing data on the clusters does not provide more security [30]. So, Hadoop system also lacks of providing sufficient security facilities. The following are the identified security issues of Hadoop cluster [31]:

- An unauthorized user may access an HDFS file via the remote procedure call (RPC) or via HTTP protocols and execute arbitrary code or carry out further attacks
- A data block of a file at a DataNode may be read/written by an unauthorized client via the pipeline streaming data-transfer protocol.
- An unauthorized client may gain access privileges and submit a job to a queue or delete or change priority of the job.
- An unauthorized user may eavesdrop/sniff to data packets sent by Datanodes to client.

MapReduce: *a programming model for cluster system*
MapReduce [32] is a programming model compatible for cluster architecture. In fact, the model provides an interface

for the distribution of sub-tasks and gathering of outputs. So, it has capability to process large data sets within a distributed cluster. In particular, it consists of two main primitives, namely Map() and Reduce() that are commonly used in functional programming paradigm. Both the functions are performed by the Master node in the cluster. In Map() step, the Master node splits a problem into number of sub-problems, each for one slave node. On the other hand, each slave processes the assigned sub-problem and passes the computed result to the Master. Finally, the Master integrates all the results (sent by the respective slave nodes) to get the final answer. Obviously, the entire process is similar to the *scatter* and *reduce* strategy followed in Message Passing System. It is important to note that a slave node may invoke the Map() procedure again for further dividing the problem. The key contributions of the MapReduce are scalability and fault-tolerance achieved when processing massive data on a large cluster. It greatly simplifies the task of writing a large-scale analysis on distributed data for many types of analysis.

Healthcare big data

In general, healthcare data include medical information like patient basic information, clinical data, doctor's written notes and prescriptions. At present, medical domain is extensively using new technologies such as capturing devices, sensors, and mobile applications. More medical knowledge or discoveries are being accumulated in a constant flow. Collection of genomic information becomes cheaper. As a result, medical images such as X-Rays, CT and MRI-scan results, surgery and implants results, laboratory records, genomic information, medication information, insurance details, and other patient-related data are continuously being included into healthcare databases. Hence, the volume of healthcare database is growing exponentially. However, one of the main reasons behind such expansion is the inclusion of medical images. *For instance*, CLEF medical image data set contained around 66,000 images between 2005 and 2007 while just in the year of 2013 around 300,000 images were stored everyday [33]. Further**,** as per Seibert's report, medical image data can range anywhere from a few megabytes for a single study (e.g., histology images) to hundreds of megabytes per study (e.g., thin-slice CT studies comprising upto 2500+ scans per study) [34]. In addition, patients' social communications in digital forms are increasing day by day. For more on healthcare big data, one may refer the recent reviews [17,35–38].

Unfortunately, the explosive growing rate of complex healthcare data directs that managing healthcare data by traditional software tools (methods) and/or hardware is very difficult (or impossible). However, this vast amount of complex data yields many opportunities (e.g., quality services, reducing healthcare cost, detecting uncommon disease pattern, etc.) for us, and that can be achieved through effective analysis of data. More specifically, big data analytics has the potential to improve care, save lives, and lower costs by discovering associations and understanding patterns and trends within the data. Truly, healthcare big data itself is not useful, unless effective analysis is to be made. One may see the important article [39] in this purpose. Certainly, a novel advanced technology is essential to perform a real-time analysis over such a big data set. In particular, such an analysis must help the government to provide value-added services to the citizens.

Let us now highlight some important sources of healthcare data in the Sect. 3.1.

Sources of healthcare big data

It is true that health data are voluminous and heterogeneous. The reason is that they come from different internal and external sources that are available at multiple locations (geographic as well as different healthcare providers' sites) in numerous legacy and other applications (transaction processing applications, databases, etc.). Further, the data may be in multiple formats, e.g., flat files, .csv, relational tables, ASCII/text, etc. Some very common examples of the internal and the external sources of health data are listed below.

External sources

Web and social media data Data from specific health cites, Facebook, Twitter, LinkedIn, blogs, and the like belong to this source.
Machine to machine data This includes readings from remote sensors, meters, and other vital sign devices.

Internal sources

Biometric data It contains finger prints, genetics, handwriting, retinal scans, x-ray and other medical images, blood pressure, pulse and pulse-oximetry readings, and other similar types of data.
Human-generated data Medical data collected from Electronic Medical Records, physician's notes (paper documents) and interpretations, interviews with the patient, etc. are examples of human-generated data. These are usually unstructured or semi-structured or both.

It is important to note that all these data have to be pooled, cleansed and prepared for the purpose of analytics.

Healthcare data and 5V's

Possibly, the most widely quoted definition of big data today includes 5 Vs (characteristics) namely volume, variety, velocity, veracity and value. Interestingly, healthcare data is the prime example of big data, as it satisfies all these features. Researchers have appropriately correlated both the two. In this regard, one may refer research articles [3,4,40]. However, an explanation also is given below claiming how healthcare data are treated as big data.

Volume On the basis of the general discussion on big data, the feature 'volume' refers to scale of data**,** and the data are usually measured in Terabytes and petabytes. In healthcare database, information such as *personal information, radiology images, personal medical records, 3-D imaging, genomics, and biometric sensor readings*, etc. are being included day-to-day. Obviously, all these information collectively increase the size and complexity of the database to a great extent. To know more in favor of voluminous feature of healthcare data, one may refer the studies [33,34,41]. Clearly, the usage of systems like Hadoop, MapReduce, and MongoDB is becoming much more common with the healthcare research communities because of their capability to store and compute large volume of data [42,43].

Variety We have already known that variety relates the format of big data. In reality, the health data also are *structured, unstructured* and *semi-structured*. Example of structured information is clinical data, whereas data such as doctor notes, office medical records, paper prescriptions, images, and radiograph films are unstructured or semi-structured.

Velocity The concept of velocity for big data exactly correlates with that of healthcare data, since most of the health data are in form of paper files, X-ray films and scripts and the growing rate of such data is now dramatically increasing.

Veracity Keeping the meaning of this property in mind in context of big data, we may affirm that it signifies here the degree of *trust* about the healthcare information. In other words, veracity feature of healthcare data gives information certificate about correct diagnosis/treatment/prescription/procedure/outcomes, etc.

Value It is well accepted that the *value* (in context of big data) refers to worth of information**.** Based on this principle, the creation of value for patients should determine the *rewards* for all other actors in the system. Achieving high value for patients must become the overarching goal of healthcare delivery. Definitely, if value improves, then patients, payers, providers and suppliers all can be benefited while the economic sustainability of the healthcare system increases. Thus, this 'V' is excessively unique, as it represents the required outcomes of big data processing.

Opportunities with healthcare big data

Due to digitization and interconnection of healthcare data, significant benefits (opportunities) are achieved today. The potential advantages include *quality administration*, *reduction of workload*, *savings* of *consultation time*, detecting diseases at *earlier* stages to treat it more*easily* and *effectively* with *reduced cost*, detecting *healthcare fraud* (that involves the filing of dishonest healthcare claims) more quickly and efficiently, managing particular individual and population health properly, etc. Some of the major benefits (mainly achieved through analytics) are detailed below as much as possible for showing more practical insights.

Benefits to patients Healthcare data can assist patients in making right decision at right time. In fact, analytics of patient data does this job. Further, analytics may be applied to identify the individuals who need "*proactive care*" or *changes* in their lifestyle to avoid degradation of health condition. Thus, it results in improving the health of patients while decreasing the cost of care. A concrete example in this respect is the Virginia health system Carillion Clinic project, which uses predictive models for early interventions [44].

Benefits to researchers and developers (*R and D*) Patient data collected from different sources help *research and development* to improve quality of research about new diseases and therapies. Actually, R and D may propose new algorithms (especially related to data mining and machine learning) to detect new diseases that may cause epidemics. In this respect, one may refer the studies [45,46].

Benefits to healthcare providers Healthcare data assist the providers to frame preventive acts. Further, the providers can design new strategies to take care for patients. Accordingly, it reduces the number of unnecessary hospitalizations.

Clinical operations The health data set is capable to provide comparative effectiveness research to decide more practical and clinically important approaches. It also suggests the cost-effective ways to diagnose and treat patients.

Public health On analysing disease patterns, tracking disease outbreak and its transmission ensures to improve public health-surveillance and speed-response. Example includes *faster development of more accurately targeted vaccines*, e.g., *choosing the annual influenza strains.* In this context, Lazer et al. state that turning large amount of data into actionable information can be used to identify the needs, especially for the benefit of populations [46]. In addition, it provides services, predicts and prevents crises for the individuals.

Genomic analytics It assists to execute gene sequencing more efficiently and cost effectively. Ohlhorst states that genomic analysis must be a part of the regular medical-care decision process and the growing patient medical record [47].

Detecting spreading diseases earlier Healthcare analysis has ability of early prediction of viral diseases before their spreading. Surely, this may not be possible by analysing the *social logs of the patients* suffering from a disease in a specific geo-location [48]. After all, analysis helps the healthcare *professionals* to advise the victims by taking essential preventive measures.

Fraud detection Misuse of a person's medical identity to wrongfully obtain healthcare goods, services, or funds may be detected from healthcare analysis. Undoubtedly, fraud in medical claims can increase the burden on the society. Importantly, predictive models like decision tree, neural networks, linear regression, etc. can be used to predict and prevent fraud at the point of transactions [49].

Evidence-based medicine It involves the use of statistical studies and quantified research by doctors to perform diagnosis. This practice enables doctors to make decisions not only based on their own perceptions but also from the best available evidences. It is, indeed, an effective advantage obtained from healthcare data.

Secondary usage of health data The secondary usage of health data deals with aggregation of clinical data from finance, patient care, administrative records to find valuable insights like identification of patients with rare disease, therapy choices, clinical performance measurement, etc.

Issues of healthcare data

Despite various benefits of healthcare data, some key issues confronting the healthcare services are increasing day to day. These are as follows:

Aging population
Significant number of medical errors
Uneasy access of healthcare information
Inefficient operation of large data
Demand for quality and safe healthcare services, as the resources (e.g., number of doctors, hospitals, laboratories, etc.) remain at the same level.

Challenges in healthcare data

We have been already acquainted with the potential opportunities and issues of healthcare data. On analysing the healthcare issues, a possible list of *challenges* (also research opportunities) is presented below to achieve the aforementioned opportunities.

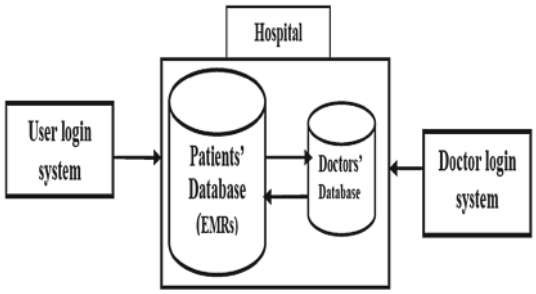

Fig. 1 Block-diagram of simple e-healthcare system

Ease of understanding and use Understanding clinical notes (especially unstructured papers) in right context and its appropriate use are really the challenging tasks.

Scalability Operating efficiently large amount of medical data (especially image data) and extracting potentially useful information from the data in order to *reduce medical errors* are crucial jobs. In fact, these two jobs remind us the scalability issue. Accordingly, developing an appropriate model for supporting immediate response of user query is a complex task.

Cost Analysing *genomic* data itself is a computationally intensive task. Now, combining such data with standard clinical data adds extra layers of complexity.

Capturing patient *behavioral data* through several sensors and analysing these data are no doubt the big questions. However, the key challenge in context of healthcare system is the *security issue* (explained below).

Security issue In Healthcare Information System (HIS), security should be the top*priority* from day one. At any cost, patient data (especially *sensitive data*) should be protected by adopting comprehensive physical security, data encryption, user authentication and the latest standard-setting security practices and certifications. In fact, such an issue mainly arises due to the use of cloud (i.e., distributed) computing architecture in HIS because cloud hosts the patient information and provides different services to the authorized users. So, we should pay attention at different levels of health system to impart security in healthcare data

A conceptual e-health system and its need

An e-health system is a system dedicated for healthcare services supported by electronic processes and it was started

around in 1999. Although there exist several definitions in this regard, it is well-accepted that the adoption of e-health system using ICT tools can achieve significant improvements in quality and safety of healthcare delivery. Such a system primarily aims to improve access of clinical data, safety of patients and efficiency of healthcare process. Also, it is capable to reduce clinical errors and healthcare cost. A very generalized block-diagram of a simple conceptual e-healthcare system is first presented in Fig. 1 to get some basic ideas on e-health system. The primary goal of this architecture is to provide an *interface* between application and the users belonging to a hospital for easy access of patient data. The architecture comprises of four basic modules, namely *user login* system, *patient data base*, *doctors' database* and *doctor's login* system for a hospital. The modules are briefly explained below.

User login module It examines authentication of user (*i.e.*, patient). Two parameters, namely *user-name* and *password* are essential in this purpose. Obviously, if an user is authentic, then he/she is permitted to access his/her record stored in patient database.

Doctor's login module This module is responsible to authenticate the doctors belonging to the hospital. The similar approach as adopted for *user-login* module is to be followed here.

Patient database It is, indeed, the *digitized* version of patient information and known as electronic medical record(EMR). In particular, EMR contains medical and treatment history of patients, e.g., patient name, Address, Mobile number, Mail-id, Date of birth (DOB), medical information, surgery results, side effects, referred Doctor information, etc. The stored information for each patient can be visualized as a single file, named as EMR file, and it is usually managed by hospital. A sample EMR file is shown in Appendix A.

Surely, EMR has several *advantages* over paper records—such as it allows clinicians to track data over time, to identify easily which patients are due for preventive screenings or check-ups, to check how their patients are doing on certain parameters like blood pressure readings or vaccinations and to monitor and improve overall quality of care within the practice. One may note that some important information of patient (called as *sensitive data*) may be blocked to access (by applying appropriate encryption schemes) for the purpose of security. However, authentic doctors may be allowed to access patient information including the secret information using the decryption scheme. *Doctors' database*: it primarily maintains the doctors' profiles, containing name, mail-id, qualification, area of treatment, experience, assigned patients, etc. of each doctor.

Table 2 Comparison between EMR and EHR

Sl. no.	EMR	EHR
(i)	It is a database maintained by a CDO	It is the aggregation of EMRs
(ii)	It is owned by one CDO. So, it does not share with other CDOs	It may be owned by several CDOs. So, it shares with multiple CDOs
(iii)	EMRs are supplied by hospitals, vendors, clinics, etc	EHRs are run by community, state or national organization

Distributed e-healthcare system

We have been already acquainted with the primary job of any e-health system. Now, to fulfil the job to a great extent, sharing information among different hospitals is necessary. More specifically, medical notes, medication information, medical test results and allergy information about patients, specialized doctors in hospitals, infrastructure of hospitals, etc. should be shared from one hospital to another hospital when and where they need it. Surely, operating healthcare data (consisting of structured, un-structured and semi-structured data) and sharing the data over several locations may not be managed by the traditional DBMSs generally used in standalone system. Also, their storage capacity is not enough to store the so-called big data. That is why we may think about centralized database system for healthcare data to tackle the issues of traditional DBMSs. However, any centralized database system still faces several issues like the presence of single control point, bottleneck problem, etc. but these may be resolved using distributed database system. After all, the expected requirements of healthcare system may not be satisfied, unless the researchers pay attention in designing well-organised distributed e-health system.

Drawbacks of EMRs in distributed healthcare system and the solution

If distributed healthcare system is adopted for quality health services, then a major disadvantage to EMR is that the records cannot be easily and accurately shared among the users in distributed system. The reason is, information about the patients in EMRs is collected from several providers (healthcare units) like hospital, pathological labs., radiology, pharmacy, etc, i.e., it may not be fully complete during sharing information. That is why EMRs of several sites are integrated by an entity named as electronic health record (EHR) (explained below), and the EHR is then shared by Care Delivery Organizations (it is also detailed below). The main differences between EMR and HER are summarized in Table 2:

- *Electronic health record* (EHR) It is a repository of information regarding the health status of patient, in computer processable form. In fact, an EHR is supposed to contain

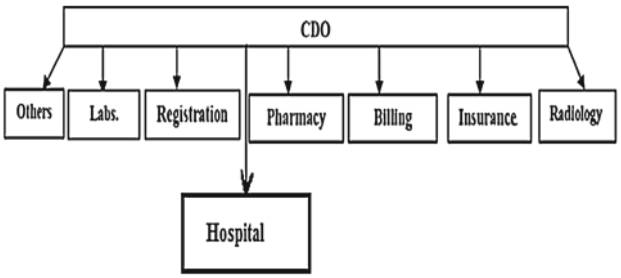

Fig. 2 A simple block diagram of CDO

all the necessary information about the patients collected from several providers plus evidence-based tools to make intelligent decisions. Thus, EHR maintains the total health of the patients. Note that it may store data in structured or unstructured or in both format.

- *Care delivery organizations* (CDOs) These are actually the low-level entities directly connected to the end-users. The primary mission of these entities is to deliver the healthcare-related products and services. Each organization comprises of a few healthcare units such as hospital, radiology, laboratory, pharmacy, billing and so forth. The coordination among the CDOs in the cloud is performed by means of EHR unit, and it empowers information sharing and interoperability. However, interchange between various CDOs offices and the EHR is made conceivable by utilizing HL7 convention. A simple block diagram of CDO is shown in Fig. 2.

The health information system (HIS) framework proposed by Ahmed [15]

The proposed framework for HIS consists of the components, namely *cloud*, CDO, EMR, EHR, security module and the BDA for e-health system. All these modules (except security module) are already explained in the earlier sections. The model claims high-level integration of data, interoperability and sharing of EHRs among healthcare providers, patients and practitioners. In fact, the use of cloud aims at sharing of EHRs among authenticated users. The data analytics part of the model is assumed to analyze patient data to provide right intervention to the right patient at the right time. However, the main *limitations* of this model are as follows:

(i) It hides implementation details,
(ii) The set of security constraints adopted for preserving patient data are not powerful. In particular, the constraints are likely the security schemes used in Hadoop system and these can easily be hacked.

Proposed conceptual distributed health information system framework

Health information systems have been created by many countries and international organizations during past 10 years. However, the most important and challenging task in designing HIS is to organize and maintain patient information repositories securely, accurately and in speedy manner. It may be reminded here that most of the earlier HISs' are either confined to a particular organization or multiple organizations with a centralized system (e.g., Zookeeper). As a result, availability and accessibility of patient information are not easy, and these two barriers affect the overall performance of the system. Actually, every service is here controlled by a central server. So, if the central server fails, then the entire system fails. Further, sharing information at right time among healthcare practitioners is essential in context of diagnosis of patients. But this may not be properly satisfied through centralized system, since the inter-operation and opinion sharing in such system is very slow. To overcome these limitations, researchers have paid attention to design cloud-based distributed HIS. But the use of cloud computing technologies in e-health causes privacy and security concerns for patient data. In other words, attackers may be able to access data stored in cloud if there do not exist sufficient security mechanisms by cloud service provider (CSP). In healthcare sector, data security and privacy protection are required not only by the patients themselves, but they are also demanded by law in most countries. So, data security is treated as a key factor in healthcare system. Nkhoma [50] asserted that it must be considered during any cloud computing implementation. According to the author of the present study, less academic works emphasizing to resolve security issues for e-health distributed system are available in literature. However, there exist several studies in the literature discussing the security issues of the cloud computing but the solutions are not discussed there.

The present section deals with the proposed distributed HIS framework focusing to preserve privacy and security of patient information. The Fig. 3a gives an overall idea on the suggested framework, consisting of some essential components. Further, a schematic of EHR (an essential component of the HIS) is depicted in Fig. 3b to get better visualization of the model.

The Fig. 3b says that EHR integrates a set of EMRs (each maintained by a CDO).

Note Figure 3a shows a single EHR but this does not treat it a centralized system. The reason is that the EHR is to be accessed via cloud removing the concerns of access and memory (on the basis of the principle of cloud computing). In other words, the cloud will spread and replicate the data at several servers to achieve flexibility and reliability.

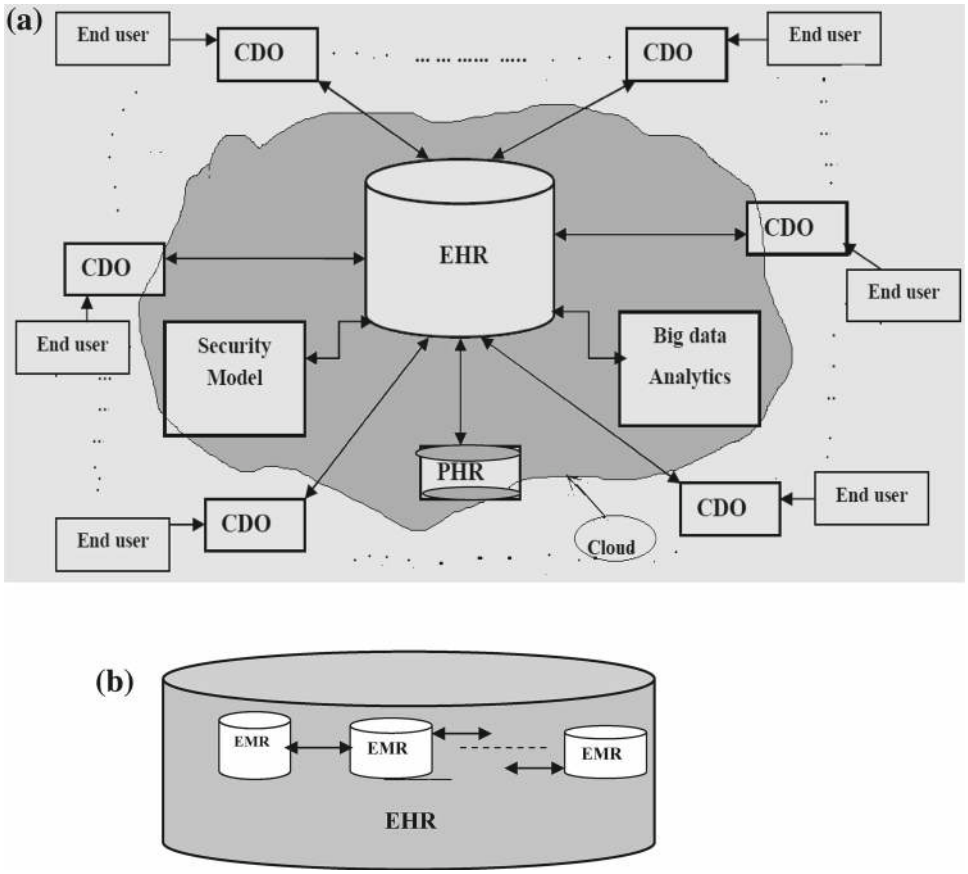

Fig. 3 **a** A proposed conceptual framework for distributed Health Information system. **b** A schematic of EHR

Major components of the proposed framework

Simply looking into the Fig. 3a, it is clear that the proposed architecture consists of some major components, namely cloud, EHR, PHR, CDO, Data analytics, End-users, etc. Each of these is discussed below. Certainly, all the components need to be connected via cloud. So, the first and the foremost component of the system is the cloud.

(i) *Cloud* On the one hand, the cloud hosts patient information and provides different services to the authorized users. On the other hand, its computing part supplies the necessary services over the network, and big data analytics analyzes lots of data to gain insights and to find the exact behaviour of data.

Why cloud computing in healthcare system?
Today, healthcare providers and insurance companies are extensively using certain kind of electronic medical record systems. However, almost all of them store medical records in centralized databases. Unfortunately, a centralized system causes big problems with respect to *accessibility* and *reliability* of stored information. The reason is that every service is here centrally controlled. Surely, if the central point fails, then the entire system fails. Further, sharing information at right time among healthcare practitioners is essential in context of diagnosis of patients. But this may not be properly satisfied through centralized system, since the inter-operation and opinion sharing in such system is very slow.

With these issues in mind, administrative boundary is translated to sharing information among EMR systems through cloud computing environment. Technically, cloud has capability to store huge amounts of data with backup facility, and its computing part provides an attractive IT platform to *cut down* the cost of electronic health record systems in terms of both ownership and IT maintenance burdens for many medical practices. In short, the cloud environment makes the records accessible to patients, practitioners and health plan services. That is why cloud computing environment is preferred here.

(i) *Public health record* (PHR) This component of HIS is linked with EMRs as well as EHR. Actually, it contains aggregated information of EHR (*i.e.*, groups of observations with summary statistics based on those observations). For this purpose, we may use some existing tools to collect, track and share past and current information about one's health. The summarized information may assist to save individual's money and inconvenience of repeating routine medical tests. More

Table 3 User's registration form

Sl. no.	Basic information
(i)	Full name of the user
(ii)	Short address
(iii)	Profession
(iv)	Preferable accessing *hospital* comprising *name*, *id*, etc. (already stored in cloud database)
(v)	Cell-phone number
(vi)	Digital signature

importantly, it is used for public health and other epidemiological purposes, research, health statistics, policy development, and health service management.

(ii) *Care delivery organizations* (CDOs) It may be kindly reminded that the structure and the role of each CDO are already described in Sect. 4.1 with the help of a block diagram (as shown in Fig. 2). Recall that each CDO maintains EMRs, and the information in EMRs is collected from several units such as hospital, radiology, laboratory, pharmacy, billing and so forth. Further, end users are connected to cloud through CDOs across the country, whereas the CDOs are connected to EHR.

(iii) *Data analytics* It is, indeed, big data analytics (BDA). Truly, different types of big data require different analysis methods. For the present system, one may refer a conceptual healthcare analytic model as pictured in Fig. 6 (placed in Appendix A). However, in order to analyze multi-Terabyte EHR databases in cloud, the BDA part in the present system may opt existing significant distributed architecture Hadoop based on programming models like MapReduce, NoSQL, etc., or some new designed approaches based on High Performance Computing (HPC) that may touch most or all of the data.

(iv) *End users* The end users are here usually the patients, doctors, nurses, specialists, technicians, researchers or other individuals, or groups. Surely, as several users are assumed to be connected in cloud, so *privacy* and *security* of data are here the matter of concern. Although research on various security issues surrounding healthcare information systems has been heated over the past few years. In particular, ISO/TS 18308 standard [14] is the most popular one for EHRs. However, these schemes are not sufficient to secure health data while accessing in cloud. That is why some more steps are taken here to tackle this crucial job. These are explained below.

First, any new user must*register* into the cloud providing the following mandatory information other than*user-login* (name) and *password* before accessing cloud data. The information are noted in Table 3.

Fig. 4 A sketch of m-digit hospital-id

Hospital-id The hospital-id mentioned in Table 3, it is a unique semi-auto-generated number (say, an m-digit number) whose first n digits denote the *location* of the hospital (linked in cloud) and the rest $(m - n)$ digits for other information such as *department, unit*, etc. A sketch of an *m-digit* hospital-id is depicted in Fig. 4.

Note It may happen that there may have number of hospitals in a location connected via cloud. If so, then few digits, say k digits (out of n digits), may be reserved to represent serial numbers of the hospitals depending on the year of establishment.

Suggested some others ids' or codes for securing cloud-based health system.

Based on $m-$digit *hospital-id*, the following necessary *ids*' are generated.

Patient-id It is a unique semi-auto-generated M-digit (M> m) number whose first 'n'digits match with the first n-digits of *hospital-id* (one may refer the sample file of Appendix A for verifying the *patient-id* as well as the *hospital-id*). Obviously, the rest (M-n) digits are auto-generated digits.

User-id A semi-auto-generated *unique number* of M digits (M>m) is assigned to each *new user* after successful registration into cloud. The number satisfies the following:

- The first$n$ digits of the number must match with the first $n$digits of the *hospital-id* that he/she *prefers* to access. Obviously, the rest (M-n) digits are auto-generated digits.
- The *user-id* need not necessarily be same as the *patient-id*, but its size is assumed to be same as *patient-id*. The user may be allowed to access other hospitals connected in cloud.

Patient-code It is a unique 13-character alphanumeric code that is generated after filling the patient detail form (as shown through EMR file placed in Appendix A). The characters (i.e., places) in the generated code represent the following:

- The first two places of the code denote the first two characters of the *city* at which the hospital is located, e.g. RA for Ranchi.
- Next two places take the first two characters of the *state* at which the hospital is situated, e.g., JH for Jharkhand.
- The fifth and the sixth places say, respectively, the first characters of the *first* name and the *surname* of the patient, e.g., AP for A. Pal.
- Immediate next place indicates *gender*: M or F.

- Two places next to *gender* represent the short form of the disease that attacked the patient at the time of first hospitalization, e.g., YF for yellow fever.
- Finally, date of birth of the patient (in DDMMYY format) is stored in the last four places, e.g., 021092 for 02/10/1992.

Definitely, the code is a secret code, and the healthcare authority may supply it to the patient and the doctor(s).

Now, as the research focuses much on data security, so it is necessary to discuss the basic security principles to be followed by EHR of any e-health system. The discussion is given below.

The basic security principles adopted by any EHR

Any cloud-based EHR should adopt the following common security principles for secure access:

(i) All electronic medical records, be it PHR or EHR or EMR, should be guarded through ownership-controlled encryption, enabling secure storage, transmission and access.
(ii) The creation and maintenance of EHR should preserve not only content authenticity but also data integrity and customizable patient privacy throughout the EHR integration process.
(iii) Accessing and sharing of EHR via cloud should be secured and systematic through powerful security model, since it may severely cause privacy and security problem. Recall that one major challenge to healthcare cloud is the security threats that include tampering or leakage of *patient sensitive data* in the cloud, loss of privacy of patient information and the unauthorized use of such sensitive information. Thus, a powerful and organized security module is essential for any distributed healthcare system to prevent these threats.

An overall idea on the proposed HIS

A sketch of the proposed HIS is depicted in Fig. 5a for better understanding of its working procedure. The system consists of three important parts, namely Part 1, 2 and 3. The expected role of each part is detailed below.

Part-1 This part is responsible for storing digitized data in EHR. The EHR is to be linked with PHR as well as EMR via cloud, where EMR contains the digitized version of patient data and PHR contains the aggregated data of EHR.

Part-2 It deals with *categorization* of patient information (stored in EHR) on the basis of *sensitivity* levels. This is, indeed, the *pre-processing* phase of *data analytics* for health database and it is to be performed by expertise persons. However, an automated system may be developed for this purpose. In the present system, three tiers (i.e., levels) of sensitivity are chosen for healthcare data to protect patient information:

Tier-0 (Super sensitive) Examples include disease-name (e.g., HIV/AIDS) and its status, mental status of the patient, mediclaim number,biometric identifiers). Surely, top-level security is essential to access such data pieces.
Tier-1 (Medium sensitive) Examples include date of birth, name, doctor's name, person type such as celebrities, political figures, etc.)
Tier-2 (Least sensitive) e.g., Zipcode/Pincode, Gender, blood group, name of surgery). These are, in fact, commonly accessible date.

Part-3 This part mainly cares about preserving privacy and security of patient data while accessing the data over cloud by the end users because distributed file systems like Hadoop do not provide much security facilities to escape even from unauthorized users. In particular, this part employs two levels of security (Level-1 and 2) for performing the job(starting from user authentication checking to access sensitive data), and it is named here as 2-Level Security Model (i.e., 2-LSM). A detailed diagram of Part-3 (2-LSM) is shown in Fig. 5b for better understanding the model.

Figure 5b says that a user must be authenticated at the *entry point* into cloud prior to Level-1 security. In fact, this verification must be done by the authentication process followed in the cloud. A suggested authentication process for cloud users is presented below.

Authentication and technologies

In a healthcare cloud, information offered by Contracted Service Providers (CSP: a form of cloud service provider) and identities of end users (e.g., practitioners, patients and others) should be verified at the entry level, using *user-login* and *password* (assigned to users by CSPs). For this purpose, technologies such as digital signatures, timestamps, confirmation receipt, encryption, etc. may often be preferred to establish authenticity and non-repudiation for patients, practitioners and others. Next is the *authorization* checking of the user. It is, indeed, the Level-1 security of the proposed system (discussed below).

The proposed 2-LSM (part-3 of the proposed HIS)

Recall that this part is mainly responsible for preserving privacy and security of patient data from malicious hackers, while accessing those over cloud. It consists of two levels of security, namely Level-1 and 2 for carrying out this job.

An important note Before accessing patient information, a *valid user* in the cloud may request to get a temporary *user-id* (for faster access in the cloud) if he/she is a patient too. In

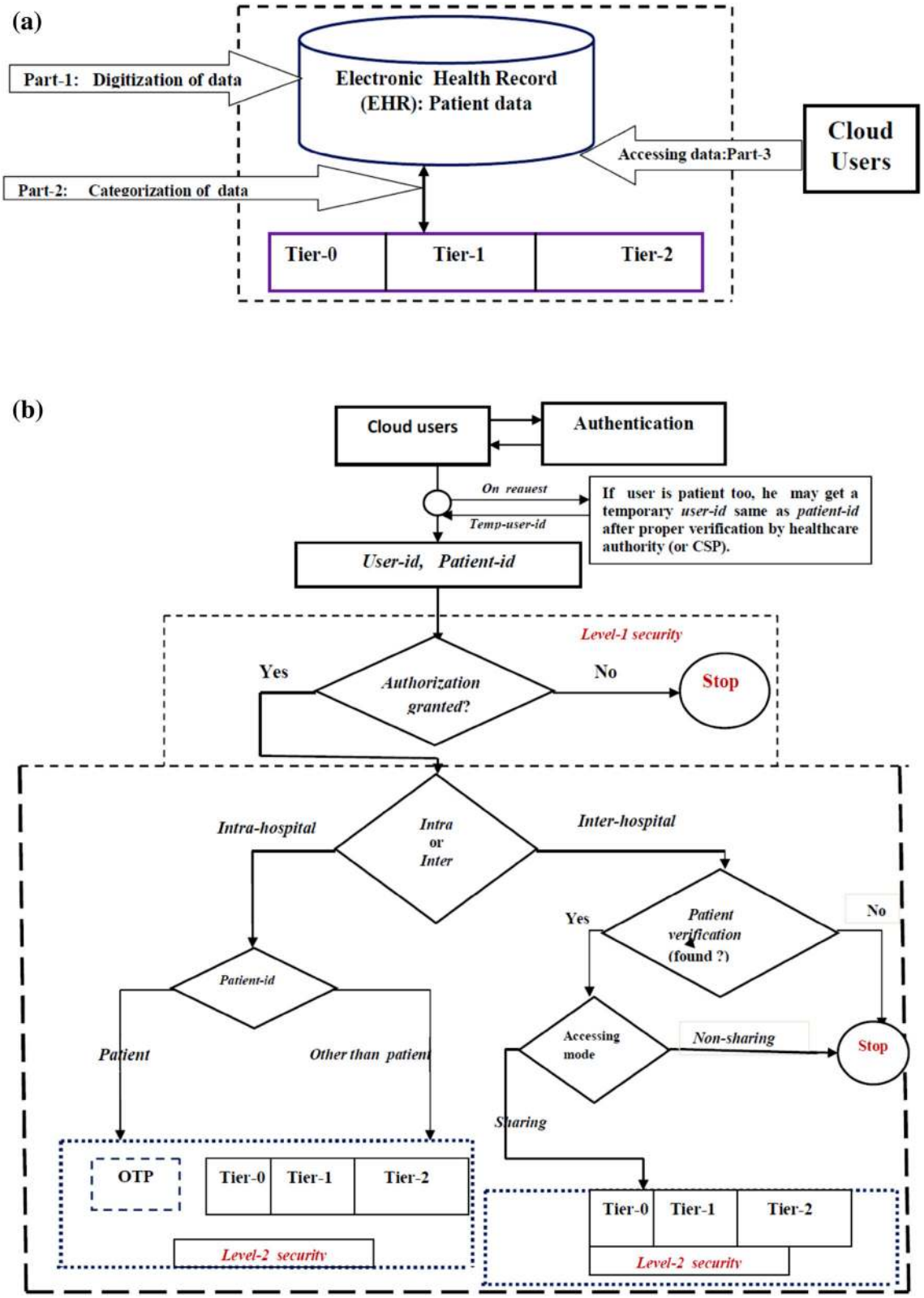

Fig. 5 **a** Diagram showing an overall structure of the proposed system. **b** Flow graph of 2-LSM (a secured healthcare big data architecture)

such a case, the user must supply his/her *user-id* as well as the *patient-id* to the healthcare authority connected in cloud. The user may be authenticated as the *user* as well as the *patient* on receiving *patient-id* as a *temporary-user-id* after sufficient verification by the authority.

Level-1 security of 2-LSM It relates to authorization of user for accessing patient data in a hospital (after authentication of user through login-process is over).

Authorization It primarily aims at controlling access priorities, permissions and resource ownerships of the users in client-server architecture. Each client is granted privileges based on his account. In healthcare cloud, users may be *blocked* to access patient information through the privileges granted by the healthcare authority. In particular, it is to be decided using *patient-id*. Actually, healthcare authority must play an important role in this respect. A short note in this regard is given below.

- *Role of healthcare authority* The healthcare authority will set up *teams* (groups) of approved practitioners to access patient EHRs. The authority is also accountable for providing privileges to each team on the basis of their specialization, disease kind and complexness, and the role played for treatment.

Structure of group (*or team*)

- The group may be of doctors, nurses, specialists, lab. technicians, pharmacists and other practitioners like researchers.
- Group members are usually to be selected from different hospitals, cities or regions.
- The healthcare authority is supposed to assign every group with digital signatures to sign the medical certificates provided by them. Further, it is assumed that individual member in a group will possess completely different privileges according to his/her job and profession.

Encryption techniques like AES [51] and RC4 (Ron Rivest-of RSA Security in 1987) and authentication techniques such as One Time Password (OTP) and Two Factor Authentication (2FA) are some well-accepted protection schemes from unauthorized access of data, and these also may be employed here. Besides, the proposed system must follow the confidentiality rule defined by International Organization Standards (ISO) in ISO-17799.

As per the concept of Level-1 security, a user is primarily supposed to access patient data if he/she is authorized by the authority.

Level-2 security of 2-LSM
This level of security will be applied at each of the suggested two cases, namely *intra-hospital* (i.e., patient admitted hospital) and *inter-hospital* (i.e., other than the patient admitted hospital but linked in cloud) as explained below. Actually, on the basis of *user-id* and *patient-id* (as supplied after successful entry into the cloud), we may identify the case of *intra* or *inter* hospitality.
Identifying user under intra-hospital or inter-hospital:

Input to the cloud *patient-id* and *user-id*
If the first *m* digits of *user-id* match with the first *m* digits of *patient-id* of the patient, then it is to be assumed that the user falls under *intra-hospital*, else he/she is in *inter-hospital*.

Case-1 Intra-hospital

(i) *User as the patient* If the *user* is *patient* himself, then he/she receives an OTP. After entering the OTP into the system, the patient is to be allowed to access his or her data.
(ii) *User other than the patient* The user may be permitted to access different levels of patient sensitive data as follows:

For Tier -2 data access, user supplies the OTP if received. Note that the user will receive an OTP if and only if he/she is authentic and permitted by authorized persons to access such type of data.
For Tier-1 data access, the user supplies the OTP (if received) and a *Token*. The concept of Token is very similar to *CAPTCHA* that assists one to access a specific resource.
For Tier-0 (super sensitive) data access, user supplies in sequence the OTP sent to him/her for accessing data, a *Token* and the *patient-code*.

Case-2 Inter-hospital

If the *patient-id* supplied by the user is *valid* (*i.e.*, the patient belongs to a valid hospital linked to cloud), then the followings are to be examined:

- It verifies whether *data accessing* mode is sharing or non-sharing. If data sharing is *yes* (*i.e.*, granted), then go for Tier-1, 2 and 3 accesses (as discussed earlier), else *stop*.

For better understanding the 2-LSM and implementation details of Tiers-0,1 and 2, one may refer Appendix-B.
Note We may facilitate *private* cloud for storing Tier-0 and 1 data, whereas *public* cloud for storing Tier-1 data for faster access.

Discussion and analysis of the proposed framework

In this section, discussion about implementations and analysis of the proposed framework is presented.

Implementation The suggested semi-automated *ids*' such as *hospital-id*, *patient-id*, *user-id* and *patient-code* are implemented in Java-1.4.1 on a stand-alone Pentium-4 m/c running on Mandrake Linux OS 9.1 and tested for several cases. The two-level security model (2-LSM) introduced in the present study is also implemented in Java-1.4.1 on the same stand-alone m/c and verified. It is successfully working on the stand alone m/c but yet to apply in cloud. Regarding the perfect implementing of the proposed system, we may opt existing distributed framework Hadoop and its compatible programming tools like MapReduce or NoSQL, or any cloud-based HPC to touch most or all the data.

Analysis Some highlighting points of the proposed system comparing with the model suggested by Ahmed [15] are presented below.

Implementation of the introduced cloud-based e-health system is *easier*, as high-level description of the model is presented. However, the architecture proposed by Ahmed [15] hides implementation details.
The present framework emphasizes more on preserving privacy and security of the patient data, as Hadoop-based file system has several security issues for unauthorized users. But the model proposed by Ahmed [15] does not pay much attention in this respect.
The present model provides facilities for *intra-hospital* and *inter-hospital* cases of patients for faster retrieval of data using patient-id and user-id (supplied by the user just after successful entry into the cloud). However, the Ahmed's model did not do so.
The idea on *patient-code* in the study is a unique feature for preserving security of patient data.

In addition, the authors hope that the proposed framework on implementation will attain the following positive characteristics:

Incremental growth Adapting new CDOs (connected to more users) as well as servers (to support more services) is easy, and it is one of the salient features of any distributed system.
Utilization of numerous information Easing the patients to utilize a mix of enormous information, supporting distributed computing technologies.
Right decision at right time Offering opportunity to take preferences of huge measures of healthcare information, and prescription of proper mediation to the patient at right time.
Support of mobile computing Capability to combine big data with mobile and cloud computing.

Managerial implications of the proposed model The proposed model concentrates not only on protecting patient data but also it has potentiality to attain the following managerial implications:

Integrating hospitals across the country linked via cloud (based on the proposed model) claims reduction of healthcare cost and workload.
Availability of skilled doctors, nurses, staffs and other necessary infrastructures must provide faster relief to the patients through quality treatment.
It is hoped that hospitals with EHR and clinical decision support system (CDSS) will deliver prompt diagnosis with reduced consultation and treatment time.
In addition, the system may assist the government to monitor whether the hospitals are setup as per the *norms* setup by medical council. In other words, periodical check-up may assist government in taking necessary measures against disqualifying hospitals.

In summary, the use of the model based on EHR may significantly upgrade consistency, predictability of healthcare services.

Conclusion and future scope

In summary, big data analytics is recognised as a multidisciplinary information processing system in the areas like business, government, media, education and healthcare. In particular, it is a growing area with the potential to provide useful insight in healthcare.Effective integration of data mining and medical informatics and its subsequent analysis using big data techniques will no doubt impact healthcare delivery-cost and improved healthcare results via well-informed decision making. More specifically, big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Importantly, the big concern in context of big data is the privacy and security.

The first contribution of this research is the provision of an overall picture on big data and healthcare data for non-expert readers. The other one is the adoption of a holistic view to build an organized healthcare model for protecting patient data. The model provides high-level integration and sharing of EHRs. The suggested framework applies a set of security constraints and access control that guarantee integrity, confidentiality and privacy for medical data.
Future scope The present architecture aims to fulfil the following targets after successful implementation on the basis of availability of the necessary infrastructures:

Provision of high-quality but low-cost medical services for the patients through interaction among the practitioners across the country and BDA as well.
Rapid and widespread use of big data analytics across the healthcare organization and the healthcare industry.
Search engines and social networks can help to gather people's reactions and monitor the conditions of epidemic diseases.

Appendix A

Some beneficial examples of e-healthcare system **(reported by IBM in 2013:** http://www03.ibm.com/industries/ca/en/healthcare/documents/Data_driven_healthcare_organizations_use_big_data_analytics_for_big_gains.pdf).

Example 1 Premier (an U.S. healthcare alliance network) has more than 2700 members, hospitals and health systems, 90,000 non-acute facilities and 400,000 physicians. The network has assembled a large database of clinical, financial, patient and supply chain data, with which the network has generated comprehensive and comparable clinical outcome measures, resource utilization reports and transaction level cost. These outputs have informed decision-making and improved the healthcare processes at approximately 330 hospitals, saving an estimated 29,000 lives and reducing healthcare spending by nearly $7 billion.

Example 2 In Medical Centre of Columbia University, big data analytics in healthcare is able to perform analysis of "complex correlations" of streams of physiological data related to patients with brain injuries. This assists medical professionals with critical and timely information to aggressively treat complications. The advanced analytics is reported to diagnose serious complications as much as 48 hours sooner than previously in patients who have suffered a bleeding stroke from a ruptured brain aneurysm.

A sample EMR file

Electronic Medical Record (EMR)
Hospital id: 033356 771114
Name: Raj hospital
Place: Ranchi
City: Ranchi
State: Jharkhand
Pin/zip code: 835 288
Country: India
Patient-id: 033356 389991678
Patient name: Mr. Ankit Pal
Gender: Male
Type of person: Common
Mediclaim number: 445645898
Family member associated with the hospital: Nil
Contact no.: 9000013333
Mail-id: appal@gmail.com
Age: 24
Date of birth: 02/10/1992
Blood group: O+
Admitted date: 19/2/2016 14:48:04
Date of release: 25/2/2016 16:44:08
Assigned doctor's name: Dr. Raju Titus
Contact no.: 8877228264
Patient's diseasename: Yellow Fever
Symptoms: Heavy headache
Related details: Normal
Condition: Emergency
Allergies' by: Aspirin
Allergies' note: Normal
Name of surgery: NA
Surgery note: NA
Family doctor: Dr. Bikash
Emergency contact person: Mr. Raman
Contact no.: 9432282666
Past information about the patient: —
Next admitted hospital: —

Important note It is true that all the information stored in EMR file are *digitized*, and the information in EMR file

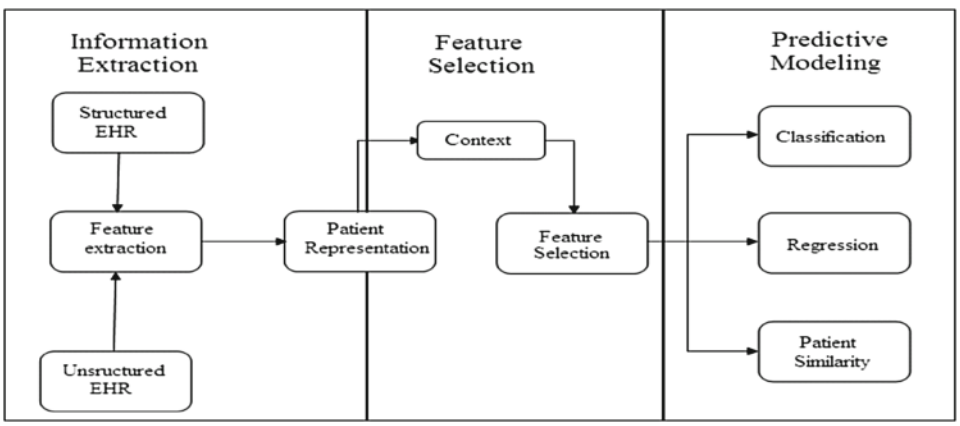

Fig. 6 A conceptual healthcare analytic model

are usually in structured form. However, the file may contain some unstructured data such as scanned copy of past record, photos, etc. Anyway, *unstructured information* such as some past or present information about the patient written in paper may be appropriately converted to structured format manually or automatically (if any tools exists) and then be entered into the EMR file. This process is known as double-reading/entry process. Further, a *unique name* of each EMR file (for each patient) may be given on the basis of his/her *patient-id* (Fig. 6).

Appendix B

Verification of user, patient and hospital-ids'

Example-1 Suppose that user-id = 033356 389991678, Hospital-id = 033356 771114 and Patient-id: 033356 389953-416. Let $n = 6$ (the first 6 digits of the numbers).

Here, the first 6 digits of the user-id is: 033356 and the first 6 digits of the patient-id is: 033356. Both are representing the first 6 digits of the hospital-id. It is, here, 033356 that matches with the first 6 digits of Hospital-id = 033356 771114 .

- Thus, user's preferred hospital and the patient's hospital are same, so it falls under *intra-hospital* case.

Example-2 User-id = 033356 389991678 and patient-id: 033356 389953416 are not same.

- Suppose that user are patient are same person, then the user, on request, may initially get a temporary user-id same as: patient-id (such as: 033356 389953416).

Example-3 User-id = 033356 389991678 and patient-id: 043456 389953416 are not same. Also, they are not the same person.

Further, 033356 $\neq$ 043456. Surely, two hospitals are not same, so it falls under inter-hospital case.

Here is a brief *explanation* about Tier-0, 1, 2.

- After logging in into the system (HIS cloud), the user (assuming that the user is authentic) may get option such as Tier-0,1 and 2.
- As soon as, the user (other than the patient) clicks over, say Tier-2, then the system may supply an OTP. If the CDO (corresponding to the patient) permits him to access Tier-2 data, then he gets an OTP to access the patient basic information. He then enters it into the system to access the Tier-2 information.
- Likewise, if the user (other than the patient) clicks over, say Tier-0, then the system asks for patient-code, an OTP sent to him and a Token (CAPTCHA).
- However, if the user is patient himself, then he is supposed to allow all the information of the data after supplying the OTP sent to him.

Terminologies

Data analysis It is the process of extracting useful information or pattern from data by using data mining techniques.

Sensitive data A specific group of personal data that is understandably subject to a stricter regulation than other types of data/information. Such data must not be accessed without consent of users or authorized persons. It comprises of wide range of information like ethical or racial origin; political opinion; religious or other similar beliefs; memberships; physical or mental health details; personal life; or criminal or civil offences.

HL7 (Health Level Seven based on the concept of 7 layers of ISO/OSI model) It was established in 1987 in the USA. Its primary goal was to develop messages in consensual formats in order to facilitate a better interoperability of Hospital Information Systems.

Epidemiology It is the study and analysis of the patterns, causes, and effects of health and disease conditions of defined populations. This focuses on public health, policy decisions and evidence-based practices for identifying risk factors for disease and targets for preventive healthcare.

Contracted Service Provider (CSP) It is also known as *cloud service provider*. Usually, this entity provides information management services relating to the communication of health information.

One-time password (OTP) It is a password number (known as personal information number: PIN) which is valid for only one *login* session or transaction on a computer system or other digital devices. Its generation algorithm typically makes use of pseudo randomness or randomness, ensuring difficult to predict it by the attacker.

Semi-auto generated (or semi-random number) A unique number whose one or more part(s) is/are supplied externally but some part(s) is/are generated by m/c based on the idea of pseudo-random number. Some such numbers or codes are generated in this study, e.g. *hospital-id* to distinguish each hospital, *patient-id* to distinguish patient, *user-id* to identify each user in cloud, *patient-code*: a unique secret code for each patient.

References

1. Bakshi K (2012) Considerations for big data: architecture and approach. In: Aerospace Conference, 3–10 March 2012, IEEE. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6187357
2. Halevi G, Henk F (2012) The evolution of big data as a research and scientific topic: overview of the literature. Biometric, issue (30). https://www.researchtrends.com/wp-content/uploads/2012/09/Research_Trends_Issue30.pdf
3. Raghupathi W (2010) Data mining in health care. In: Kudyba S (ed) Healthcare informatics: improving efficiency and productivity, pp 211–223

4. Dembosky A (2012) Data prescription for better healthcare. Financ Times 2012:19–22
5. Feldman B, Martin EM, Skotnes T (2012) Big data in healthcare hype and hope. Dr. Bonnie 360. http://www.west-info.eu/files/big-data-inhealthcare.pdf
6. Fernandes L, O'Connor M, Weaver VJ (2012) Big data, bigger outcomes. AHIMA 2012:38–42
7. CCC2011c (2011) Smart health and wellbeing. Computing Community Consortium. Springer, Berlin
8. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey & Company, New York
9. Alnuem M, EL-Masri S, Youssef A, Emam A (2011) Towards integrating national electronic care records in Saudi Arabia. In: International conference on bioinformatics and computational biology, Monte Carlo Resort, Las Vegas, Nevada, USA, July 18–21
10. Kuo AM (2011) Opportunities and challenges of cloud computing to improve health care services. J Med Internet Res 13(3):e67
11. Sultan N (2014) Making use of cloud computing for healthcare provision: opportunities and challenges. Int J Inf Manag 34(2):177–184
12. Alharbi F, Atkins A, Stanier C (2016) Understanding the determinants of cloud computing adoption in Saudi healthcare organisations. Complex Intell Syst 2:155. doi:10.1007/s40747-016-0021-9
13. Peddi VB, Kuhada P, Yassine A, Pouladzadeh P, Shirmohammadia S, Shirehjini AAN (2017) An intelligent cloud-based data processing broker for mobile e-health multimedia applications. Future Gener Comput Syst 66:71–86
14. ANSI, ISO/TS 18308 health informatics-requirements for an electronic health record architecture, ISO (2003)
15. Youssef AE (2014) A framework for secure healthcare systems based on big data analytics in mobile cloud computing environment. Int J Ambient Syst Appl (IJASA) 2(2). doi:10.5121/ijasa.2014.2201
16. Sagiroglu S, Sinanc D (2013) Big data: a review. In: International conference on collaboration technologies and systems (CTS), 2013 (May 20-24), pp 42–47
17. Zaiyin L, Ping Y, Lixiao Z (2013) A sketch of big data technologies. In: Seventh international conference on proceeding of internet computing for engineering and science (ICICSE), 2013 (Sept. 20–22), pp 26–29
18. Leventhal R (2013) Trend: big data. Big data analytics: from volume to value. Healthc Inform Bus Mag Inf Commun Syst 30:12–14
19. Priyanka K, Nagarathna K (2014) A survey on big data analytics in health care. Int J Comput Sci Inf Technol 5(4):5865
20. Google Analytics. http://www.google.com/analytics/
21. Gantz J, Reinsel D (2011) Extracting value from chaos. IDC Rev 1–12
22. IHTT (2013) Transforming health care through big data strategies for leveraging big data in the health care industry. http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today
23. Mayer-Schönberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, New York
24. Akshay K, Kumar TVV (2015) Big data and analytics: issues, challenges and opportunities. Int J Data Sci (IJDS) 1(2)
25. Xiaolong Jin W, Benjamin Wah, Xueqi Cheng, Yuanzhuo Wan (2015) Significance and challenges of big data research. Big Data Res 2:59–64
26. Bughin J, Chui M, Manyika J (2010) Clouds, big data, and smart assets: ten tech-enabled business trends to watch. McKinsey Q 56(1):75–86
27. Davis CK (2014) Beyond data and analytics. Commun ACM 57:39–41
28. Shvachko K, Hairong K, Radia S, Chansler R (2010) The hadoop distributed file system. In: 2010 IEEE 26^{th} symposium on mass storage systems and technologies (MSST), pp 1–10
29. Ghemawat S, Gobioff H, Leung ST (2003) The Googlefile system. SIGOPS Oper Syst Rev 37(2003):29–43
30. Zettaset, The Big Data Security Gap: Protecting the Hadoop Cluster
31. Das D, O'Malley O, Radia S, Zhang K (2014) Adding security to apache hadoop. Hortonworks, IBM
32. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51:107–113
33. Widmer A, Schaer R, Markonis D, Müller H (2014) Gesture interaction for content-based medical image retrieval. In: Proceedings of the 4^{th} ACM international conference on multimedia retrieval. ACM, New York, pp 503–506
34. Seibert JA (2010) Modalities and data acquisition. Practical imaging informatics. Springer, New York, pp 49–66
35. Gagana HS, Thippeswamy K (2016) Healthcare system with big data analytics: a survey. Int J Modern Comput Sci Appl (IJMCSA) 4(3)
36. Groves P, Kayyali B, Knott D, Van Kuiken S (2013) The 'big data' revolution in healthcare: accelerating value and innovation. McKinsey and Company, Chennai
37. Mathew PS, Pillai AS (2015) Big data solutions in healthcare: problems and perspectives. In: International conference on proceeding of innovations in information, embedded and communication systems (ICIIECS), pp 1–6, 19-20 March 2015
38. Sun J, Reddy C (2013) Big data analytics for healthcare. In: The tutorial presentation at the SIAM international conference on data mining, Austin, TX
39. Health Inf Sci Syst (2014) Big data analytics in healthcare: promise and potential. 2(3):1–10. doi:10.1186/2047-2501-2-3
40. Burghard C (2012) Big data and analytics key to accountable care success. IDC health insights
41. Galloro V (2008) Prime numbers. Mod Healthcare 38:14–16
42. Adrián G, Francisco GE, Marcela M, Baum A, Daniel L, Fernán GB (2013) Mongodb: an open source alternative for HL7-CDA clinical documents management. In: Proceedings of the open source international conference (CISL '13), Buenos Aires, Argentina
43. Kaur K, Rani R (2015) Managing data in healthcare information systems: many models, one solution. Computer 48(3):52–59
44. Gartenberg A (2014) IBM predictive analytics to detect patients at risk for heart failure. http://www.adamgartenberg.com/gartenberg/agartenberg.nsf/dx/ibm-predictive-analytics-to-detect-patients-at-risk-for-heart-failure. Accessed 23 Jan 2017
45. Ghani KR, Zheng K, Wei JT, Friedman CP (2014) Harnessing big data for healthcare and research: are urologists ready? Eur Urol 67(3):e58
46. Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of google flu: traps in big data analysis. Science 343:1203–1205
47. Ohlhorst F (2012) Big data analytics: turning big data into big money. Wiley, USA
48. Ren Y (2011) Monitoring patients via a secure and mobile healthcare system. IEEE symposium on wireless communication
49. Konasani RV, Mukul B, Krishnan KP (2012) Healthcare fraud management using big data analytics. A whitepaper by trendwise analytics
50. Nkhoma MZ, Dang DPT (2013) Contributing factors of cloud computing adoption: a technology–organisation–environment framework approach.- Int. J Inf Syst Eng 1(1):38–49
51. United States National Institute of Standards and Technology (NIST) (2001) Advanced encryption standard (AES): federal information processing standards publication, vol 197. November 26